DOCUMENT RESUME

ED 275 695 TM 860 570

AUTHOR Eignor, Daniel R.; Stocking, Martha L.
TITLE An Investigation of Possible Causes for the Inadequacy of IRT Pre-equating.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-86-14
PUB DATE Apr 86
NOTE 55p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Equated Scores; Higher Education; *Latent Trait Theory; Mathematical Models; *Mathematics Tests; *Measurement Techniques; Research Methodology; Scaling; Standardized Tests; Statistical Studies; Test Items; *Test Theory
IDENTIFIERS Calibration; LOGIST Computer Program; *Scholastic Aptitude Test; *Three Parameter Model

ABSTRACT

A previous study of pre-equating the Scholastic Aptitude Test (SAT) using item response theory provided unacceptable equating results for SAT-mathematical data. The purpose of this study was to investigate two possible explanations for these unacceptable pre-equating results. Specifically, the calibration process, which made use of the three-parameter model and LOGIST, and the linking procedure used to place parameter estimates on the same scale were further investigated in a two stage process to see if either was responsible for the poor IRT pre-equating results found for the SAT-mathematical data in the previous study. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275695

RR-86-14

RESEARCH REPORT

AN INVESTIGATION OF POSSIBLE CAUSES FOR THE INADEQUACY OF IRT PRE-EQUATING

Daniel R. Eignor
Martha L. Stocking

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Helen C. Weidemiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Educational Testing Service
Princeton, New Jersey
April 1986

TM 860 570

2 BEST COPY AVAILABLE

An Investigation of Possible Causes for the Inadequacy of IRT Pre-equating[1,2,3]

Daniel R. Eignor[4]
Martha L. Stocking

Educational Testing Service

[1]An earlier version of this paper was presented at the annual meeting of AERA, Chicago, 1985.

[2]This study was supported by Educational Testing Service through Program Research Planning Council funding.

[3]The authors would like to acknowledge the assistance or advice of Marilyn Wingersky, Nancy Wright, Ted Blew, and Jane Lucas in performing this study.

[4]Authors' names appear in alphabetical order.

Copyright © 1986. Educational Testing Service. All rights reserved.

Abstract

A previous study of pre-equating the Scholastic Aptitude Test (SAT) using item response theory provided unacceptable equating results for SAT-mathematical data. The purpose of this study was to investigate two possible explanations for these unacceptable pre-equating results. Specifically, the calibration process, which made use of the three-parameter model and LOGIST, and the linking procedure used to place parameter estimates on the same scale were further investigated in a two stage process to see if either was responsible for the poor IRT pre-equating results found for the SAT-mathematical data in the previous study.

An Investigation of Possible Causes for the Inadequacy of IRT Pre-equating

Daniel R. Eignor
Martha L. Stocking
Educational Testing Service

Introduction

The current thrust of research devoted to the applications of item-response theory (IRT) has generated an active interest in the use of IRT methods in the solution of score equating problems (see Cook and Eignor, 1983). Because of the special properties of IRT models, users are often able to solve problems not amenable to traditional equating methods. (See Angoff, 1971, for a discussion of traditional methods.) For other situations, IRT equating offers an alternative against which to evaluate traditional methods. In addition, a number of other important outcomes accrue from the use of IRT for equating tests; among these are 1) improved equating, including better equating at the ends of the scale where important decisions are often made, 2) greater test security through less dependence on items in common with a single old form, 3) easier re-equating should items be deleted, and 4) the possible reduction of bias or drift in equating introduced when traditional methods are used over time in certain situations, most notably when the equating samples for the old and new forms are not random samples from the population.

While the above listed outcomes accrue as the result of the application of any IRT equating method, if the test forms to be equated can be pre-equated using IRT methods, a number of additional advantages result. Pre-equating refers to the process of establishing conversions from raw to scaled scores prior to the time the new test is administered operationally as an intact final form. The process depends on the adequate pretesting of a pool of items from which the new test will be built, the calibration of

these items using IRT methods, and the utilization of a linking scheme to place the IRT parameters from the pretested items on the same scale. Among the additional advantages offered by IRT pre-equating are the following: 1) since equating using IRT pre-equating methods is possible prior to the actual administration of the test, new forms can be introduced at low volume special administrations, a particular problem if traditional methods are used; 2) since pre-equating permits linkages to many old forms, it is the most likely of any equating method to yield acceptable results should testing legislation mandate the disclosure of pretest or equating items; 3) pre-equating would allow more time to do reasonableness and quality control checks, which are normally done in a hurried fashion due to score reporting deadlines; and 4) pre-equating would actually permit a reduction in the usual score reporting cycle while simultaneously allowing more time to do the equating itself. In short, the listed advantages that can potentially result from the use of IRT pre-equating build a strong case for investigating the application of this equating method to new test forms developed by large scale admissions or achievement testing programs, although, to date, only a few such investigations have taken place (Bejar and Wingersky, 1982; Eignor, 1985). In this study, some further investigations of pre-equating the Scholastic Aptitude Test (SAT) mathematical section initially described by Eignor (1985) will be reported.

General Review of The Previous Pre-equating Study

In 1983 and 1984, a large scale IRT pre-equating study of the Scholastic Aptitude Test (SAT) verbal and mathematical sections using the three parameter logistic model was conducted at Educational Testing Service (see Eignor, 1985; also Eignor and Cook, 1984). The purpose of that study was to determine the extent to which item parameters estimated on SAT-verbal and SAT-mathematical pretest data could be used for equating purposes in a situation where intact final form SAT testing data has normally been used. The items that appear in any final SAT form come from multiple pretests and to the extent that the item parameter estimates are sensitive, for instance, to the context or position in which the items appear, there may be differences between these parameter estimates and parameter estimates generated using data from the actual final form administration, resulting in a discrepancy between equating based on pretest item parameter estimates and intact final form item parameter estimates. More specifically, in the previous study, verbal and mathematical items appearing in two final SAT forms, 3ASA3 and 3BSA3, were calibrated from pretest data. Elaborate linkage systems, quite representative of the systems that would be designed were pre-equating to be considered for operational use, were utilized to get parameter estimates for the items, contained in multiple pretests, on the same scale. The two verbal sections, one from 3ASA3 and the other from 3BSA3, were both part of one linkage system and the two comparable mathematical sections were part of the other.

The effects of using the parameter estimates, obtained from the pretest data, on the equating process were evaluated in the following way. Each of the SAT-verbal and SAT-mathematical final forms under study, when administered for the first time operationally, had been equated by conventional linear methods to two different old forms and the results of the equatings averaged. These equatings were redone using item parameter estimates based on the pretest data and item parameter estimates generated from the intact final form administration. In each case, IRT true-score equating (Lord, 1980) was performed. For each form, the IRT equating based on pretest statistics was compared to the IRT equating based on intact final form data and the linear equating used operationally when each form was put on scale. IRT equating based on intact final form data and linear equating results were used as criteria in the study for the following reasons: (1) In recent IRT equating feasibility studies (Petersen, Cook, and Stocking, 1983; Kingston and Dorans 1982), it was demonstrated that intact form IRT true-score equating in a viable equating method for aptitude test data; and (2) the linear methods actually performed to put the forms on scale operationally have undergone many years of scrutiny through their use for operational score reporting purposes. Two SAT-verbal forms and two SAT-mathematical forms were used so that the consistency of results could be assessed.

The results of pre-equating the two forms of SAT-verbal, when compared to the intact final form IRT equatings, varied considerably, ranging from reasonably acceptable for Form 3ASA3 to unsatisfactory for Form 3BSA3. Contributing reasons for the inferiority of the Form 3BSA3 pre-equating results, having to do with the location of reading comprehension items at the end of pretest sections, were advanced and discussed. The verbal results reported had clear implications for

changes in test development practice, having to do with the positioning of pretest and final form reading comprehension items, if pre-equating the SAT-verbal section were to become a real possibility.

The results of pre-equating the two forms of SAT-mathematical, when compared to the relevant intact final form IRT equatings, were fairly similar to each other and had to be considered only marginally acceptable at best. Unlike the unsatisfactory pre-equating of Form 3BSA3 verbal, contributing reasons for the discrepant 3ASA3 and 3BSA3 mathematical pre-equatings could not be clearly advanced. For certain of the mathematical items demonstrating large differences in item response functions between pretest and final form, the positions of these items in the pretests could be offered as an explanation for the differences. For the other items demonstrating large differences, no explanation, other than that there appeared to be higher percentage of four-choice quantitative comparison items in this group, could be advanced.

For the three unsatisfactory pre-equatings (one verbal and two math), perhaps of greater concern than the fact that a few items stood out as being clearly more difficult in pretest than in final form (these were the items for which the differences were clearly the result of position effects), was the fact that an overwhelming percentage of the total number of items were estimated as being at least slightly more difficult. When considered collectively, these relatively slight differences in difficulty parameter estimates were clearly a contributor to the poor pre-equating results.

In an attempt to explain why the items in pretest form were estimated as being more difficult, conventional item statistics (equated deltas) were also examined. This would provide additional information on the items; perhaps they were more difficult when placed in pretests than in a final form, and the item parameter estimates are simply corroborating this fact. Mean differences in equated deltas (pretest minus intact final form) were formed for Form 3ASA3 mathematical and Form 3BSA3 mathematical. For 3ASA3, the difference was .36, while for 3BSA3, the difference was .01. Hence, for 3ASA3, the conventional item data provided consistent results with what was observed in studying the item parameter estimates, but for 3BSA3, the results were not at all consistent.

In conclusion, Eignor (1985) was unable to explain why the items in pretest form were estimated as being more difficult than in final form, or provide an explanation for the unsatisfactory pre-equating results, particularly for SAT-mathematical Form 3BSA3, but did offer some suggestions, The purpose of the present study is to attempt to isolate certain of the factors that may have actually caused the poor SAT-mathematical pre-equating results, and to attempt to improve upon these results.

Particulars of Previous Study Relevant to Current Study

Two particular design features of the previous study have relevance for the study described in this paper. First, the data design for the SAT-mathematical data in the previous study included a chain of 14 three parameter logistic IRT item calibrations, each of which involved a separate LOGIST (Wingersky, et al, 1982, Wingersky, 1983) calibration run. Scattered throughout these calibrations were the pretest administrations of the items that later composed the intact final operational forms of SAT-mathematical designated 3ASA3 and 3BSA3. Superimposed on each calibration run was a linking/scaling procedure (Stocking and Lord, 1983) which, by making use of common items between adjacent calibration runs, allowed the placement of all parameter estimates on a common scale.

The calibration system from the previous study, which made use of pretest, final form, and equating section data, is reproduced in Figure 1. The SAT-mathematical final forms are actually two sections that together contain a total of 60 four- and five- choice items (35 items in one section, 25 items in the other section). The total is comprised of 40 five-choice regular mathematics items and 20 four-choice quantitative comparison items. The mathematical common item equating sections each contain 25 regular mathematics items and are built to be as parallel as possible to the 25 item SAT-mathematical section, which also contains regular mathematics items. The mathematical pretest sections contain either 35 or 25 items and are built to be as parallel as possible to the comparable length SAT-mathematical sections. Each box in Figure 1 represents a separate calibration (computer run). The dotted-line boxes within the larger boxes indicate the overlapping items that were used to place parameter estimates

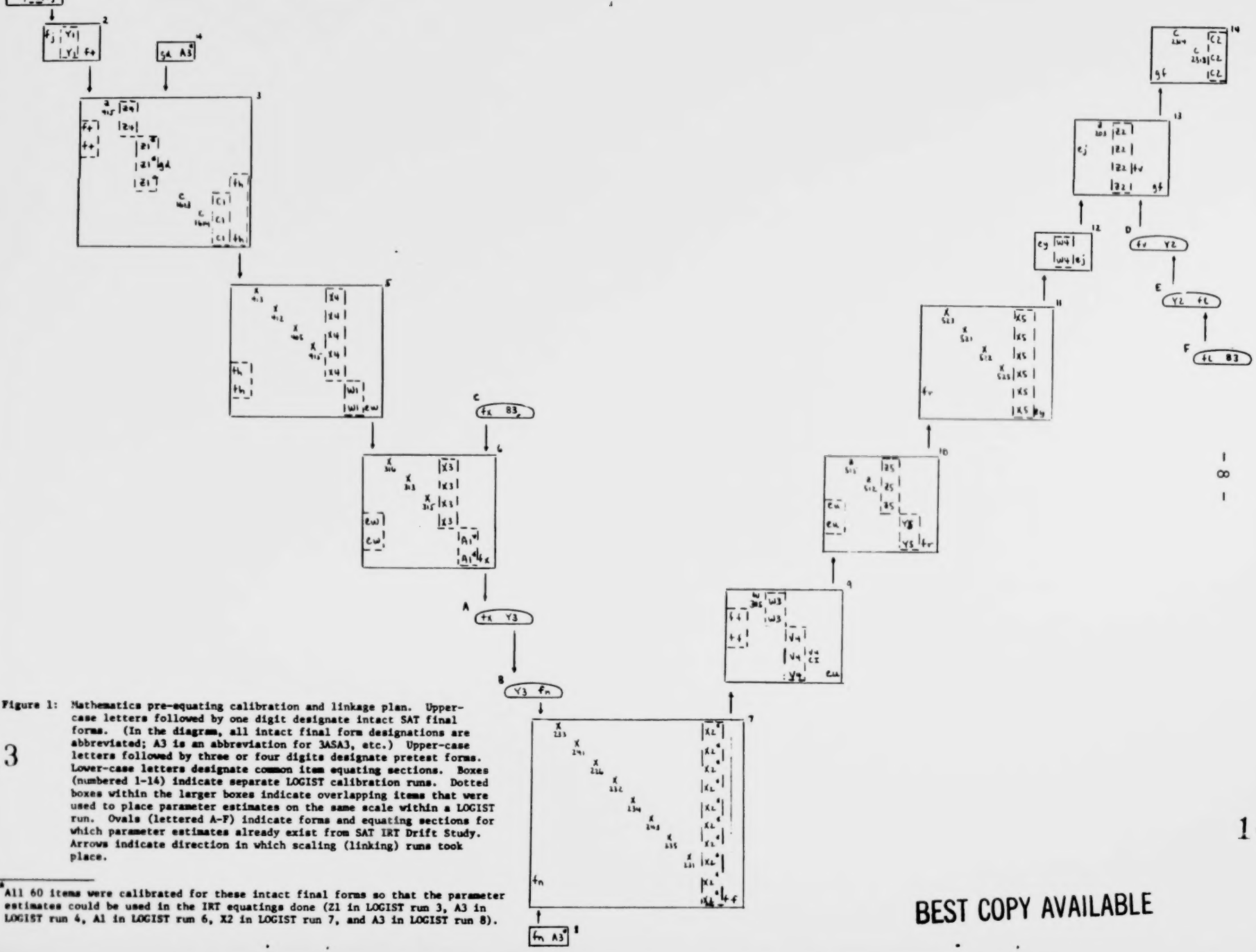


Figure 1: Mathematics pre-equating calibration and linkage plan. Upper-case letters followed by one digit designate intact SAT final forms. (In the diagram, all intact final form designations are abbreviated; A3 is an abbreviation for 3ASA3, etc.) Upper-case letters followed by three or four digits designate pretest forms. Lower-case letters designate common item equating sections. Boxes (numbered 1-14) indicate separate LOGIST calibration runs. Dotted boxes within the larger boxes indicate overlapping items that were used to place parameter estimates on the same scale within a LOGIST run. Ovals (lettered A-F) indicate forms and equating sections for which parameter estimates already exist from SAT IRT Drift Study. Arrows indicate direction in which scaling (linking) runs took place.

[a] All 60 items were calibrated for these intact final forms so that the parameter estimates could be used in the IRT equatings done (Z1 in LOGIST run 3, A3 in LOGIST run 4, A1 in LOGIST run 6, X2 in LOGIST run 7, and A3 in LOGIST run 8).

BEST COPY AVAILABLE

on the same scale within a single calibration run. The directional arrows between the boxes indicate the direction in which the scaling program (Stocking and Lord, 1983) was run to place parameter estimates from the separate calibration runs on the same scale. LOGIST calibration run 14 in Figure 1 was chosen as the base form for scaling purposes because it contains an SAT-mathematical form and equating section that are in common with a partial pre-calibration system recently devised (Cook, et al, 1985). The samples used for calibration purposes took either the two mathematical sections and one of the mathematical common item equating sections or the two mathematical sections and one of the mathematical pretests. Responses from randomly selected samples of approximately 3000 examinees taking each pretest-final form combination and approximately 2700 taking each final form-equating section combination were used in the calibrations.

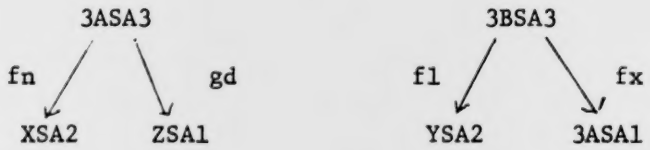
It should be noted that in the calibration process, all items contained in each 25 item equating section appearing in Figure 1 were calibrated; however this was not the case for all items in each pretest or final form. In order to reduce calibration costs, only the 35 item sections of SAT-mathematical final forms used for linking purposes (all 60 items were calibrated for the final forms actually used in the equatings) and only the 120[1] (60 items X 2 forms) mathematical pretest items which eventually appeared in final forms 3ASA3 and 3BSA3 were calibrated. Table 1, from the

---

[1]Mathematical pretest data did not exist for two of the 60 items in Form 3ASA3. Therefore, final form data had to be used for calibration purposes for one of these items and data on the other item as it appeared in an equating section had to be used.

previous study, contains the total number of mathematical items and also the total number of examinees responding to each of the 14 SAT-mathematical calibration runs. Table 2 lists the number of mathematical pretest items calibrated in each of the runs.

The following diagram depicts the equatings done operationally for 3ASA3 and 3BSA3 and the common item equating sections used.



As mentioned earlier, these equatings were redone using item parameter estimates based on the pretest items which constitute 3ASA3 and 3BSA3 mathematical and item parameter estimates generated from the intact final form administrations of 3ASA3 and 3BSA3. The final form mathematical item parameter estimates for 3ASA3, 3BSA3, and the old forms to which they were equated were placed on the same scale, which is essential for IRT equating, by being linked into the overall calibration and linking plan shown in Figure 1. This represents the second design feature of the previous study relevant to the current study. The final form parameter estimates for 3ASA3 were introduced both in calibration runs 4 and 8 in Figure 1, for 3BSA3 both in ovals C and F (i.e., calibrated items from a previous SAT scale drift study (Petersen, Cook, and Stocking, 1983) were used), for XSA2 in calibration run 7, for ZSA1 in calibration run 3, for YSA2 in oval E, and finally, for 3ASA1 in calibration run 6. For each form (3ASA3 and 3BSA3), the IRT equating based on pretest statistics

Table 1

Total Number of Items and Total Number of Examinees for each of the SAT-math LOGIST Calibration Runs

LOGIST Calibration[1] Run Number	Total Number of Items Calibrated	Number of Pretest Items Calibrated	Number of Equating Section Items Calibrated	Number of SAT-math Section Items Calibrated	Total Numbers of Examinees
1	61	1	25	35	5,441
2	85	-	50	35	4,692
3	239	35	75	129	22,071
4	85	-	25	60	2,773
5	125	4	50	70	19,007
6	151	6	50	95	16,195
7	128	19	49	60	25,291
8	84	-	24	60	2,744
9	121	1	50	69	13,735
10	127	7	50	70	13,281
11	92	7	50	35	16,594
12	85	-	50	35	5,432
13	110	1	75	35	7,838
14	97	37	25	35	7,981
	1,590	118[2]	648	823	163,075

[1] LOGIST run number refers to identification scheme in Figure 1.

[2] Pretest data did not exist for two of the 60 items in 3ASA3. Final form data had to be used for calibration purposes for one of these items and data on the other item as it appeared in an equating section had to be used.

Table 2

Number of Items Calibrated from each SAT-math Pretest Form

Pretest Form	LOGIST[1] Run No.	Total No. of Items Calibrated	No. of Items in 3ASA3	No. of Items in 3BSA3	Pretest Form	LOGIST[1] Run No.	Total No. of Items Calibrated	No. of Items in 3ASA3	No. of Items in 3BSA3
W503	1	1	-	1	X234	7	3	3	-
Z415	3	1	-	1	X243	7	4	4	-
C1613	3	18	10	8	X235	7	1	-	1
C1614	3	16	7	9	X231	7	1	-	1
X413	5	1	-	1	W305	9	1	-	1
X412	5	2	-	2	Z515	10	3	1	2
X415	5	1	1	-	Z512	10	4	3	1
X316	6	2	2	-	X523	11	3	-	3
X313	6	2	-	2	X521	11	2	-	2
X315	6	2	-	2	X522	11	1	-	1
X233	7	4	3	1	X525	11	1	-	1
X241	7	2	2	-	Z203	13	1	-	1
X226	7	1	1	-	C2314	14	21	10	11
X232	7	3	2	1	C2318	14	16	9	7
					Totals		118[2]	58[2]	60

[1]LOGIST run number refers to the identification scheme in Figure 1.

[2]Pretest data did not exist for two of the 60 items in 3ASA3. Final form data had to be used for calibration purposes for one of these items and data on the other item as it appeared in an equating section had to be used. Thus, only 58 (of 60) pretest items were calibrated for 3ASA3 and 118 (of 120) for both forms.

was then compared to the IRT equating based on intact final form data and the linear equating results used to put the forms on scale operationally.

Purpose of Current Study

The purpose of the current study is to determine whether the calibration procedure, which made use of LOGIST, or the linking procedure (Stocking and Lord, 1983), or neither of these, is the cause for the poor pre-equating results in the previous study. This can be accomplished in a two step process.

The intact final form equatings in the previous study were done using forms that were separated by only a single link in the design. That is, form 3ASA3 was equated to old form XSA2 using parameter estimates from calibration runs 8 and 7 in Figure 1 and to old form ZSA1 using parameter estimates from calibration runs 4 and 3. In a like fashion, form 3BSA3 was equated to old form YSA2 using parameter estimates from ovals F and E in Figure 1, and to old form 3ASA1 using parameter estimates from oval C and calibration run 6. It is possible, however, to perform these same equatings using parameter estimates that are separated by several links in Figure 1. For instance, 3ASA3 can be equated to old form XSA2 using parameter estimates from calibration runs 4 and 7 and to old form ZSA1 using parameter estimates from calibration runs 8 and 3. In the first equating, the parameter estimates would be separated by six links and in the second, by five links. In phase one of the investigation, the intact final form equatings were redone using parameter estimates separated by several links. If these new (multiple link) intact final form equating results then agree with the one link results, both the calibration procedure and the linking procedure were successful and there would be no need to search further for

inadequacies in either. The poor pre-equating results from the previous study must have been caused by other factors that would require investigation. However, if the new multiple link results do not agree with the one link results, then the calibration procedure and the linking procedure would need to be tested separately.

The effects of the linking procedure can be removed by running all data in one large LOGIST calibration run, with additional internal cross-links. As mentioned earlier, the design of the previous study was such that the first block of items, calibrated in LOGIST run 1 depicted in Figure 1, was connected to the last block of items by only a single chain of some 15 separate links. Each link involved LOGIST estimation and then the superimposed scaling or linking run. Any weakness in a particular link will be carried across all additional following links. A better design would have been the placement of bridging cross-links that would have strengthened the overall linkages necessary in Figure 1. Cost considerations precluded the location and calibration of these cross-links in the previous study; also, the scaling procedure used in the previous study does not provide a mechanism for simultaneously placing parameter estimates on a scale determined by multiple forms, so it is difficult to see how strengthening cross-links could have been utilized. This is not so, however, if the data is run in one large calibration run. The Eignor (1985) pre-equatings can then be repeated, and if the new pre-equating yields acceptable results, the IRT calibration process will be vindicated. The poor pre-equating results in the previous study must have been the result of the linking procedure itself

or the lack of cross-links. Individual links from that study can then be studied to find which are at fault and, perhaps, some remedy devised. If, however, the large LOGIST run does not yield acceptable pre-equating results, it must be concluded that something specific is occurring in the pretest data or in the calibration process used in this and the previous study that is causing pretest parameter estimates to be disparate from final form parameter estimates and that the three parameter logistic model, as implemented by LOGIST, can not successfully handle the specific SAT-mathematical data used in the studies.

Because the major concern in the second phase of this study has to do with the possible effects of the scaling or linking procedure on the pre-equating results in the previous study, the intact final form IRT equating to be used in evaluating the current pre-equating results should also be void of any possible effects due to linking parameter estimates from the new and old forms. This is not the case for the single and multiple link intact final form equatings examined in the first phase; the scaling procedure (Stocking and Lord, 1983) had to be used in both cases to place parameter estimates on a common scale. Because of concern about the possible effects of this scaling procedure, the intact final form IRT equatings to be used to evaluate the pre-equating results were redone in phase two, using a procedure called "b-less" equating (Stocking, 1981), which is described in the methodology section. This equating procedure is not dependent on the prior use of a parameter scaling procedure to put parameter estimates for forms to be equated on a common scale.

Methodology

LOGIST Calibration Design

As mentioned in the previous section, part of the investigation of the poor pre-equating results from the previous study involved running all data in one large LOGIST run, with additional internal cross-links. Perhaps the easiest way to pictorially represent this large run is to simply add the additional cross-links to Figure 1; this has been done in Figure 2. The previous LOGIST calibration runs that the new cross-links connect are joined to the cross-links by double-stemmed arrows in Figure 2. Common item sections that provided data for the scaling runs in the previous design now provide the overlapping items necessary for this concurrent calibration design. (See Cook and Eignor, 1983, for a general description of the concurrent calibration design.) With the addition of the cross-links, an additional 215 items were calibrated (1600 in total[1]) and an additional 38,940 abilities were estimated (202,015 in total) using the procedure described in the next section.

Item Calibration

The three parameter logistic model item parameters and examinee abilities for this study were calibrated using the program LOGIST (Wingersky, Barton and Lord, 1982; Wingersky, 1983). The estimates are obtained by a modified maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974).

---

[1]Certain items calibrated in the previous design (see Figure 1), but not necessary in the current calibration design, were deleted from this calibration. For instance, items in calibration run 8 Figure 1 were not included in the large calibration run because the 3ASA3 parameter estimates were not essential to the process of placing the pretest parameter estimates on a common scale. Hence, the total number of items calibrated in the large LOGIST run is not the sum of the items calibrated in the previous study (1590 items) and the additional cross-link items (215 items).

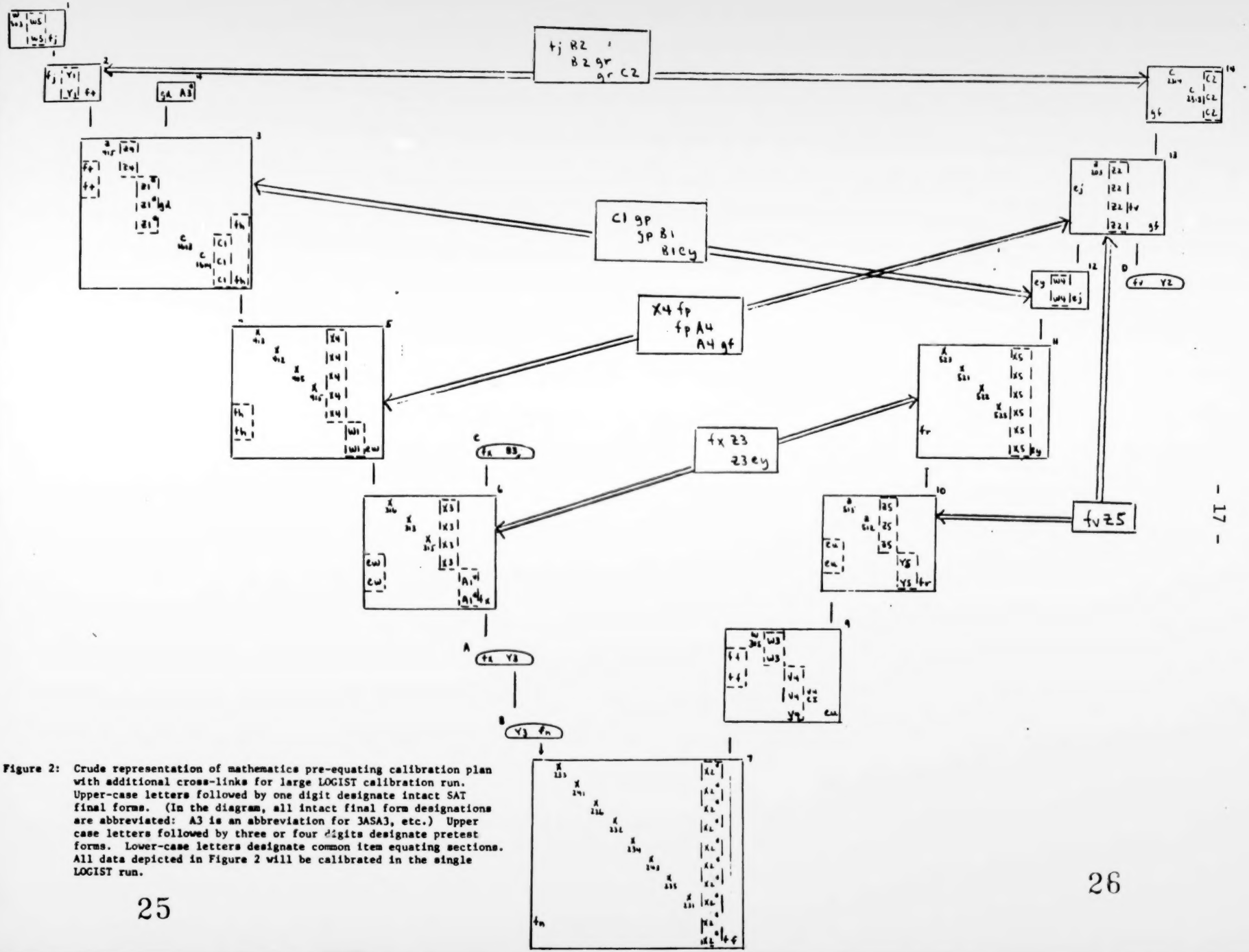


Figure 2: Crude representation of mathematics pre-equating calibration plan with additional cross-links for large LOGIST calibration run. Upper-case letters followed by one digit designate intact SAT final forms. (In the diagram, all intact final form designations are abbreviated: A3 is an abbreviation for 3ASA3, etc.) Upper case letters followed by three or four digits designate pretest forms. Lower-case letters designate common item equating sections. All data depicted in Figure 2 will be calibrated in the single LOGIST run.

LOGIST requires as input the responses to a set of items from a group of examinees, coded to reflect items answered correctly, incorrectly, omitted, and not reached. In the large concurrent LOGIST run, all items not taken by a particular sample of examinees were simply coded as not reached. In addition, the user may specify certain restrictions on the data and parameters in order to speed convergence of the iterative procedure. The major restrictions specified for the large LOGIST computer run were:

1. examinees who answered less than 15 items were not used,
2. a's were restricted to a range of .01 to 1.75,
3. c's were restricted to a range of .0 to the lesser of .50 or .75 times the proportion correct for the item, and
4. θ's were restricted to a range of -7.0 to 5.0.

LOGIST produces as output estimates of the a, b, and c for each item, and θ for each examinee.

This LOGIST calibration was the largest ever attempted: 1600 items and over 200,000 examinees. Based on the authors' previous experience, calibrations that have an item by people data matrix such as this one, where there are few cross-links, converge more slowly than calibrations with stronger cross-links. This is due, in part, to the number of stages required for changes in one block of items to be reflected in all other blocks of items. In order to minimize this effect, the final scaled difficulties from the previous calibration design were used as initial values for the item difficulties in the calibration run. Even with these initial values, this calibration took over 25 CPU hours on an IBM 3083.

IRT Equating

Although there are a number of equating techniques possible when using IRT, only true formula score equating was used in this study (Lord, 1980). The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, ξ, we have

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i + 1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \tag{1}$$

where n is the number of items in the test, $P_i(\theta)$ in the three-parameter item response function, and (k_i+1) in the number of choices for item i. If we have two tests measuring the same ability θ, then true formula scores ξ and η from the two tests are related by the equations

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i + 1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right]$$

$$\eta = \sum_{j=1}^{m} \left[\frac{(k_j + 1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] \tag{2}$$

Clearly, for a particular θ corresponding true scores ξ and η have identical meaning. They are said to be equated.

Because true formula scores below the chance score level are undefined for the three-parameter logistic model, some method must be established to obtain a relationship between scores below the chance level on the two test forms to be equated. The approach used for this study (Lord, 1980) was to

estimate the mean (M) and standard deviation (S) of below chance level scores on the two tests to be equated via the formulas

$$M = \sum_{i=1}^{n} \left[c_i \, (c_i + 1)/k_i - 1/k_i \right] \quad , \quad \text{and}$$

$$S^2 = \sum_{i=1}^{n} (c_i - c_i^2) \, (k_i - 1)^2/k_i^2 \quad , \tag{3}$$

where n is the number of items in the test, (k_i+1) is the number of choices for item i, and c_i is the psuedo-guessing parameter for item i; and then to use these estimates to do a simple linear equating between the two sets of below chance level scores.

In practice, true score equating is carried out by substituting estimated parameters into the equations (2) and (3). Paired values of ξ and η are then computed for a series of arbitrary values of θ. Since we cannot know an examinee's true formula score, we act as if relationships (2) and (3) apply to an examinee's observed formula score.

Two further points require clarification. First, the mechanics of doing IRT true-score equating based on pretest data (pre-equating) and based on intact final form data are exactly the same. What differs are the item parameter estimates that are used to calculate $P_i(\theta)$ in equation (1). In one instance the parameters have been calibrated for the item when given in a pretest, and in the other instance, when the item was given as part of an intact final form. Second, when performing score equating to two old forms using IRT true-score equating techniques, a conversion table is generated for each new form-old form relationship and then the corresponding entries in each table are simply averaged to generate the final table.

In common applications of IRT true score equating, item parameter estimates are obtained and placed on a common (IRT) scale. The equating can then be performed between any sets of items contained in this pool of items. Since one of the purposes of this study was to investigate the possible effects of the scaling procedure on the pre-equating results from the previous study, it was considered important to have a criterion equating procedure which did not depend upon any IRT scaling method.

Such a method was applied here to obtain the criterion equatings. This method requires that the two sets of items to be equated have some items in common. To perform the equating between test 1 and test 2 which have a group of items, c, in common, requires repeated applications of the IRT equating method described earlier, as follows:

1) Test 1 is first equated to its common items, c, i.e.

   score on test 1 $\longrightarrow \theta \longrightarrow$ score on common items c.

2) The common items are identical between the two tests, consequently the output from step 1, the scores on c, are then equated to the scores on test 2:

   score on common items c $\longrightarrow \theta \longrightarrow$ score on test 2.

3) The table of scores from test 1 and scores from test 2 gives the equating between the two forms.

Note that test 1 and its common items c can be on a different (IRT) scale than test 2 and its common items, also labeled c. For this reason, this equating method is described as "b-less"; it is independent of the metric on which item difficulty, b, and examinee ability, θ, are measured.

Results

Step One Results

In step one of this study, the single link intact final form equatings from the previous study were redone using new and old form parameter estimates that were separated by several links in the previous calibration and linkage plan. Of interest is whether these new many or multiple link intact final form equatings agree with the previous single link intact final form results. If they do, then neither the calibration plan in the previous study nor the linking procedure (Stocking and Lord, 1983) applied in that study can be used as an explanation for the unsatisfactory pre-equating results. Other factors must have been responsible for the unsatisfactory pre-equatings. However, if the multiple link results do not agree with the single link results, either the calibration procedure or the linking procedure, or both, may have been responsible for the unsatisfactory pre-equatings.

Two figures (one for each new form) have been prepared to summarize the results of this phase of the study. Each of the figures contains multiple plots. Because the new forms in this study (3ASA3 and 3BSA3) were each equated to two old forms, in the figures for each of the new forms, there are plots for the single equatings back to each old forms and then the equating resulting from the averaging of the single equatings. There are two plots for each equating. The first plot compares the raw to scaled score conversion line resulting from the multiple link intact final form equating to the conversion line resulting from the single link intact final form equating. The second plot contains residuals. These residuals are

simple differences between scaled scores resulting from the multiple link equating and the single link equating for each possible formula score point. The plots use the multiple link equating result as the baseline and show differences between the single link and multiple link results across the formula score scale. Figure 3 contains the multiple link and single link results for 3ASA3 and Figure 4 contains comparable results for 3BSA3.

Of most interest in Figures 3 and 4 are the results for the single equatings, not the averages. Indeed, the residuals from the two single equatings for each form of interest, 3ASA3 and 3BSA3, are approximate mirror images of each other; thus the averages are perfect, or nearly so. This is an artifact of the study design, most easily seen by an examination of Figure 1. For example, the equating of 3ASA3 from LOGIST run 4 to XSA2 from LOGIST run 7 reflects the effects of 6 linear transformations (linkings) of item parameter estimates. The equating of 3ASA3 from LOGIST 8 to ZSA1 from LOGIST 3 reflects the effects of 6 linear transformations, 5 of which are identical to those of the XSA2 equating, except in the reverse direction.

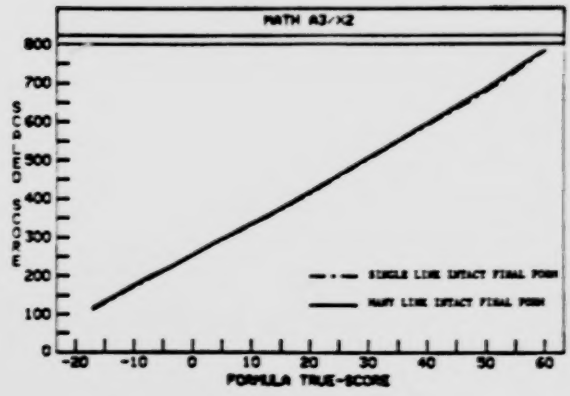
What is important in Figures 3 and 4 is the size of the discrepancies between single and multiple link equatings when equating to a single old form. These are large enough to raise the possibility that the linking procedure used in the previous study cannot be eliminated as a possible cause of the unsatisfactory pre-equatings.

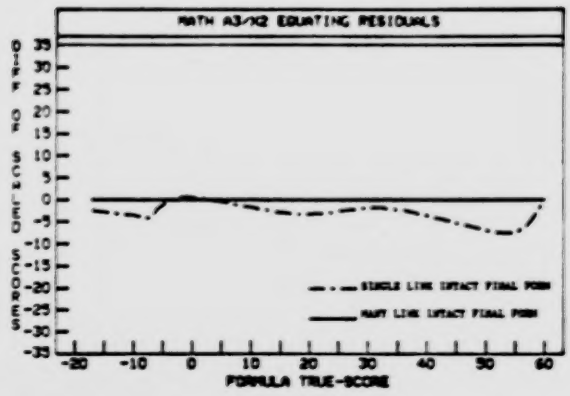
Step Two Results

In step two of this study, the data comprising the separate LOGIST runs in the previous study, along with additional cross-links, were run in

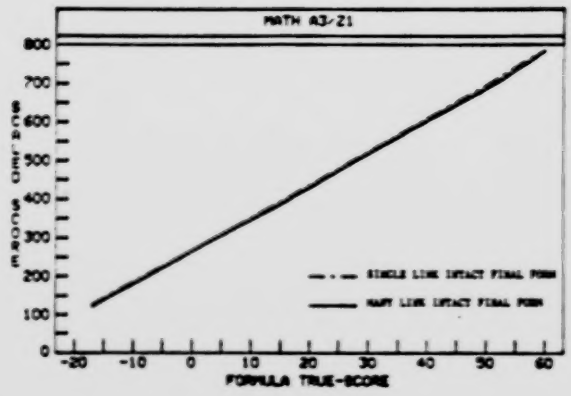
Equating Plot | Residual Plot

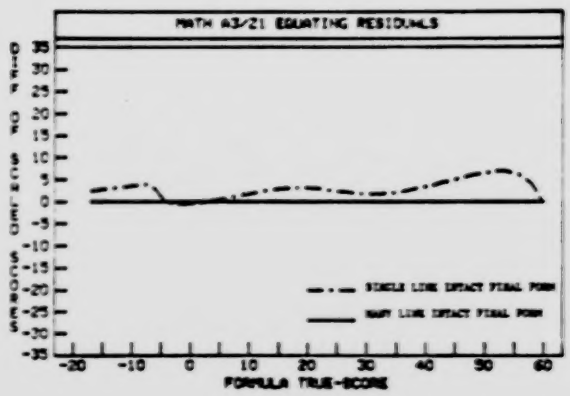
Single Equating to XSA2



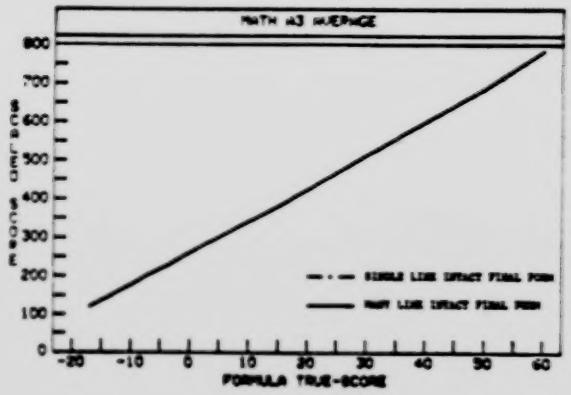


Single Equating to ZSA1





Average Equating to XSA2 and ZSA1



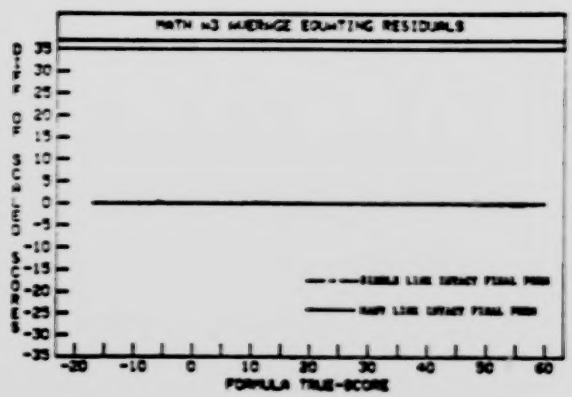
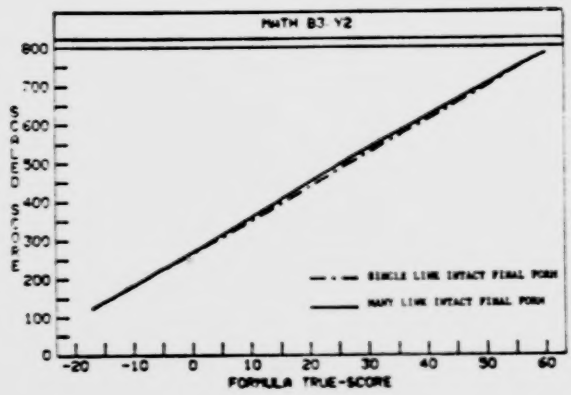

MATH A3 AVERAGE EQUATING RESIDUALS
SCALED SCORES MINUS
FORMULA TRUE-SCORE
SINGLE LINK INTACT FINAL FORM
MANY LINK INTACT FINAL FORM

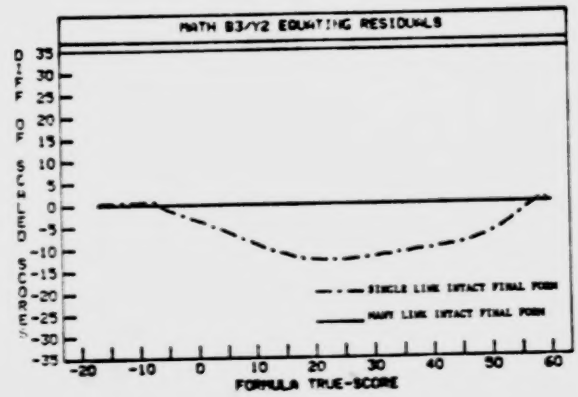
Figure 3: SAT-math Form 3ASA3 equated to SAT-math Form XSA2, Form ZSA1, and Forms XSA2 and ZSA1 - Plots of 1) single link intact final form raw to scaled transformation compared to many or multiple link intact final form raw to scaled transformation, and 2) differences between scaled scores (multiple link equating minus single link equating) resulting from the equatings.

Equating Plot

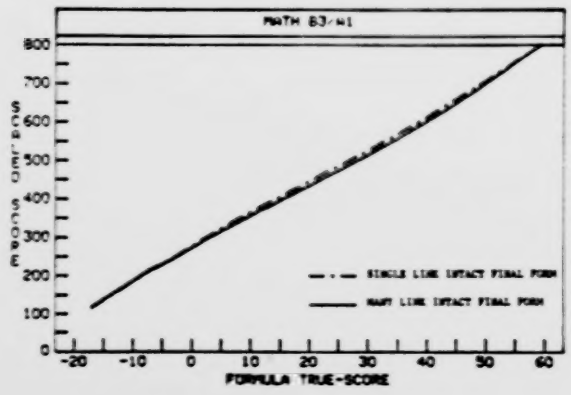
Residual Plot

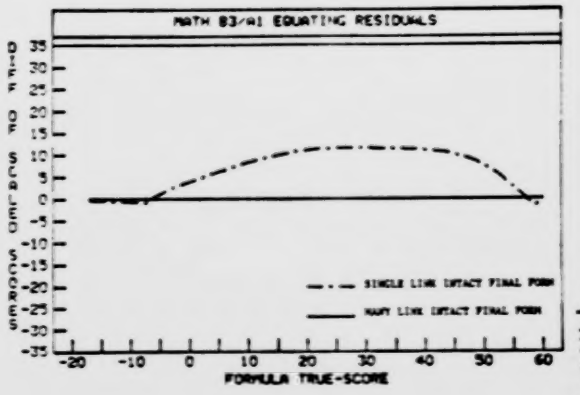
Single Equating to YSA2



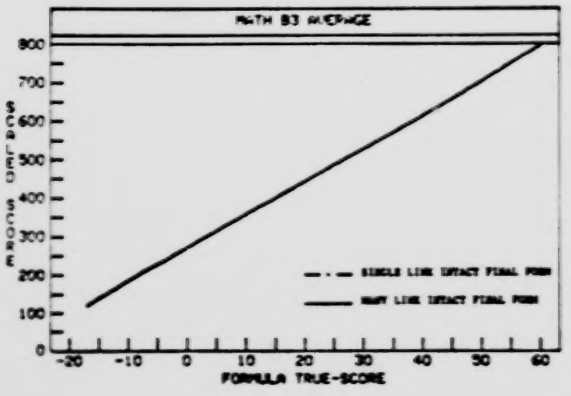


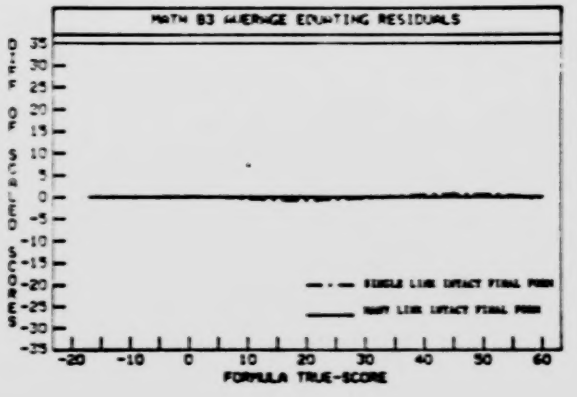
Single Equating to 3ASA1





Average Equating to YSA2 and 3ASA1





BEST COPY AVAILABLE

Figure 4: SAT-math Form 3BSA3 equated to SAT-math Form YSA2, Form 3ASA1, and Forms YSA2 and 3ASA1 - Plots of 1) single link intact final form raw to scaled transformation compared to many or multiple link intact final form raw to scaled transformation, and 2) differences between scaled scores (multiple link equating minus single link equating) resulting from the equatings.

one large LOGIST run and the IRT pre-equatings were redone. Of interest is whether the pre-equating based on this large concurrent run yields acceptable results. If so, then the poor pre-equating results from the previous study must have been the result of the linking procedure itself or the lack of cross-links. However, if the large LOGIST run does not yield acceptable pre-equating results, it would appear that something peculiar is happening in the pretest data or the calibration process that is causing pretest parameter estimates to be disparate from final form parameter estimates and that the three parameter logistic model, as implemented by LOGIST, cannot successfully handle the specific SAT-mathematical data used in this study.

Figures comparable to those prepared to summarize the results of the first phase of this study were also prepared for this phase. There are two sets of equating plots and residual plots for each of the new forms (3ASA3 and 3BSA3). The first set of plots compare the IRT pre-equating results from this study, which involved calibration of all items in a single LOGIST run, and the IRT pre-equating results from the previous study to the "b-less" intact final form IRT equating results. The second set of plots compare the two IRT pre-equating results to the intact form linear results actually used operationally to put the forms on scale. Figure 5 and 6 contain these results for Form 3ASA3 and Figures 7 and 8 contain the Form 3BSA3 results.

In addition, Table 3 contains the scaled score means and standard deviations for Forms 3ASA3 and 3BSA3 that would have resulted from use for score reporting purposes of the various equatings considered in the figures.

The means and standard deviations were computed using frequencies for the total groups taking Forms 3ASA3 and 3BSA3 at the respective initial intact form administrations.

The residual plots in Figure 5 show that the IRT pre-equating from the current study, based on the calibration of all pretest items in a single LOGIST run, provides results that are slightly more discrepant from the intact final form IRT criterion equating results than the IRT pre-equating results from the previous study, which were based on parameter estimates from multiple LOGIST runs, with parameter estimates placed on a common metric using the Stocking and Lord (1983) scaling procedure. It should be noted that the discrepancies between the average criterion and the average IRT pre-equating results from the current study are in exactly the same direction as the discrepancies for the average IRT pre-equating results from the previous study; they are just slightly more extreme through most of the raw score scale. Both IRT pre-equatings provide higher raw to scaled conversion lines than that provided by the intact final form IRT criterion equating through most of the raw score scale. The discrepancy between the current average IRT pre-equating results and the intact final form IRT criterion results is greatest around raw formula scores of 45 to 50 and, in this region, the discrepancy is between 15 and 20 scaled score points.

Using the average linear equating results actually used operationally to place Form 3ASA3 on the 200 to 800 score reporting scale as a criterion (see Figure 6), the results are quite consistent with those in Figure 5; the current average IRT pre-equating provides slightly more discrepant results,

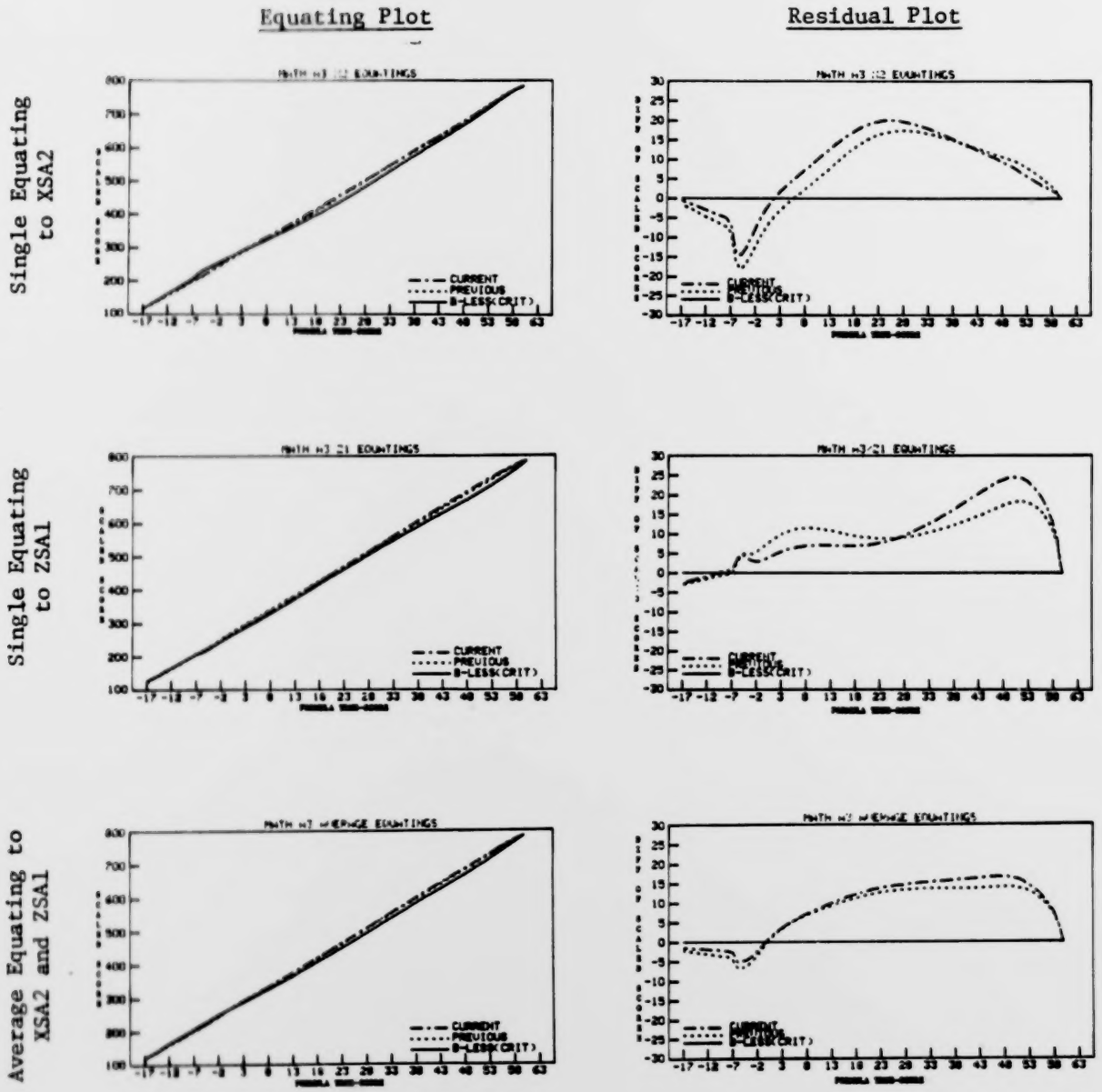
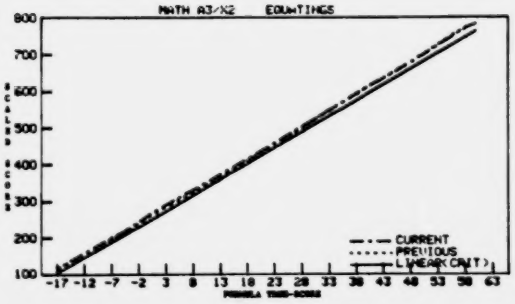
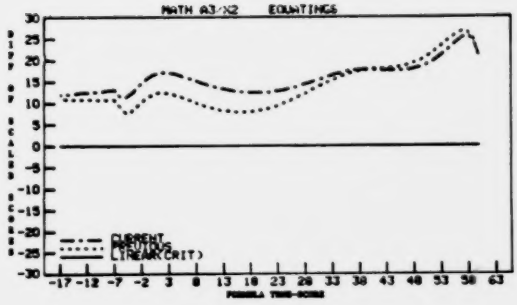


Figure 5: SAT-math Form 3ASA3 equated to SAT-math Form XSA2, Form ZSA1, and Forms XSA2 and ZSA1 - Plots of 1) previous IRT pre-equating raw to scaled transformation and current IRT pre-equating raw to scaled transformation compared to b-less intact final form IRT criterion raw to scaled transformation, and 2) differences between scaled scores (b-less intact final form IRT equating minus pre-equating) resulting from the equatings.

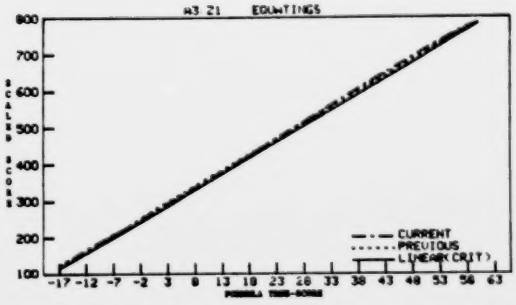
Equating Plot | Residual Plot

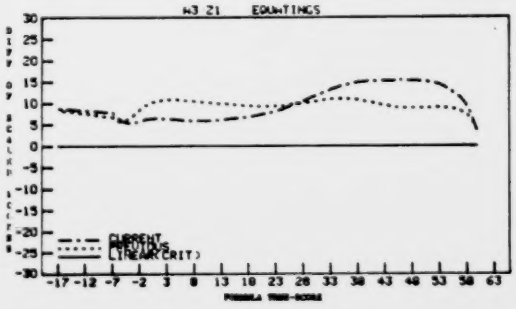
Single Equating to XSA2



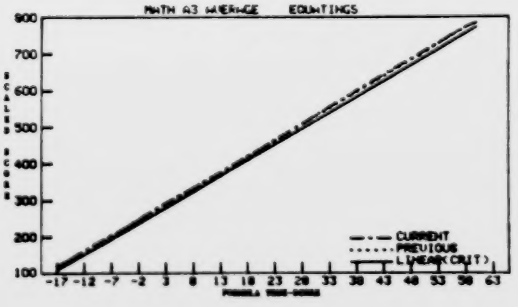


Single Equating to ZSA1





Average Equating to XSA2 and ZSA1



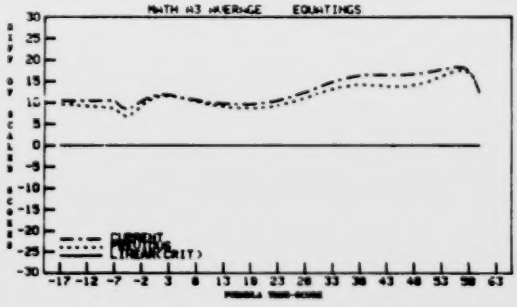
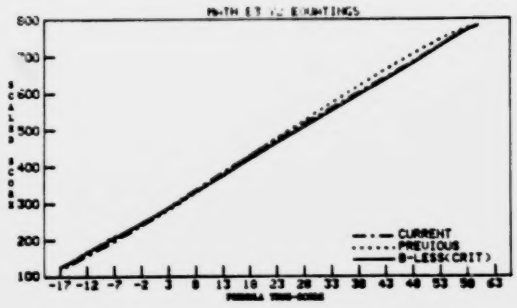
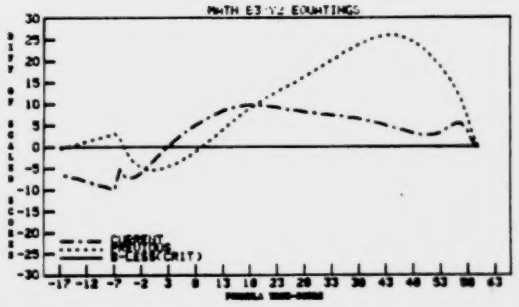


Figure 6: SAT-math Form 3ASA3 equated to SAT-math Form XSA2, Form ZSA1, and Forms XSA2 and ZSA1 - Plots of 1) previous IRT pre-equating raw to scaled transformation and current IRT pre-equating raw to scaled transformation compared to intact form linear criterion raw to scaled transformation, and 2) differences between scaled scores (intact form linear equating minus pre-equating) resulting from the equatings.

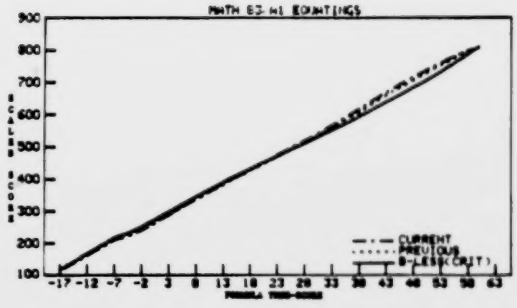
Equating Plot | Residual Plot

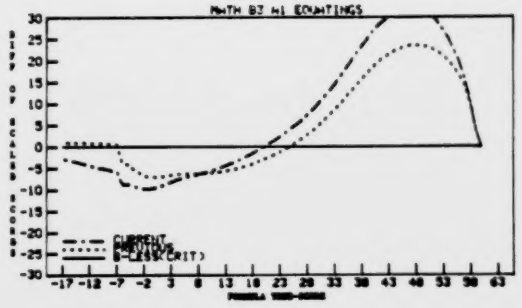
Single Equating to YSA2



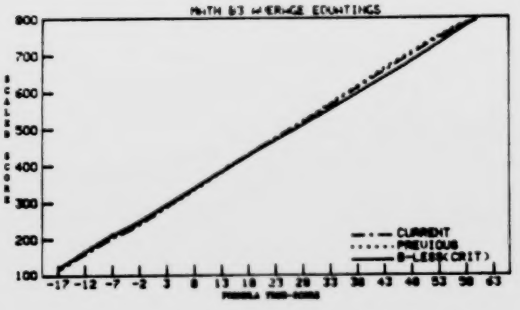


Single Equating to 3ASA1





Average Equating to YSA2 and 3ASA1



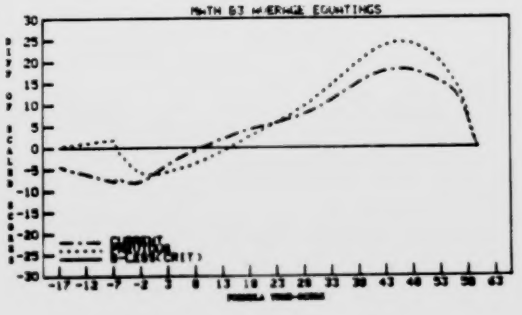
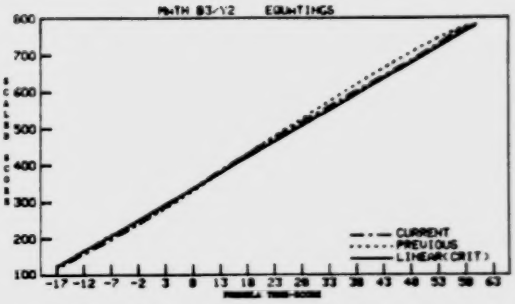


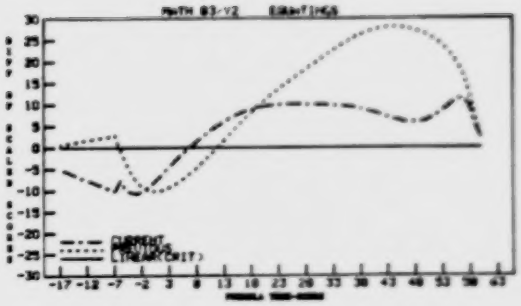
Figure 7: SAT-math Form 3BSA3 equated to SAT-math Form YSA2, Form 3ASA1, and Forms YSA2 and 3ASA1 - Plots of 1) previous IRT pre-equating raw to scaled transformation and current IRT pre-equating raw to scaled transformation compared to b-less intact final form IRT criterion raw to scaled transformation, and 2) differences between scaled scores (b-less intact final form IRT equating minus pre-equating) resulting from the equatings.

Equating Plot

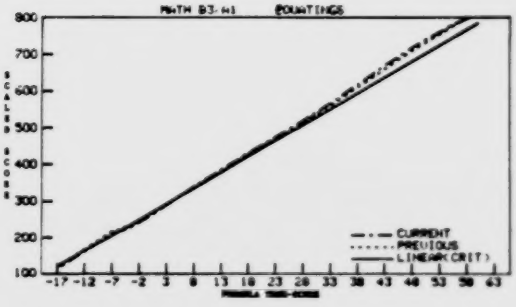
Residual Plot

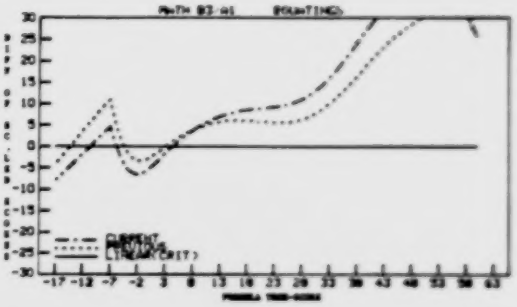
Single Equating to YSA2



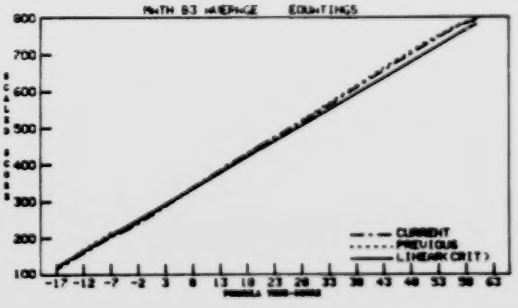


Single Equating to 3ASA1





Average Equating to YSA2 and 3ASA1



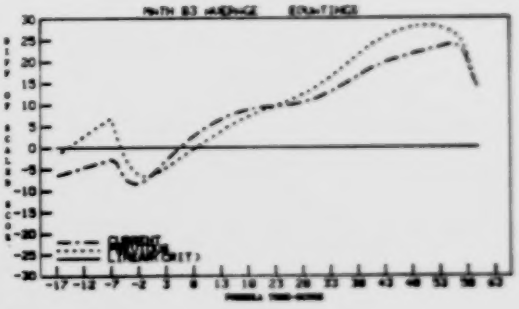


Figure 8: SAT-math Forms 3BSA3 equated to SAT-math Form YSA2, Form 3ASA1, and Forms YSA2 and 3ASA1 - Plots of 1) previous IRT pre-equating raw to scaled transformation and current IRT pre-equating raw to scaled transformation compared to intact form linear criterion raw to scaled transformation and 2) differences between scaled scores (intact form linear equating minus pre-equating) resulting from the equatings.

Table 3

Scaled Score Summary Statistics from Application of Current Study IRT Pre-equating, Previous Study IRT Pre-equating, Intact Final Form IRT Equating, and Intact Form Linear Equating Results for SAT-math Forms 3ASA3 and 3BSA3

rm	N		Current IRT Pre-equating	Previous IRT Pre-equating	Intact Final Form IRT Equating	Intact Form Linear Equating
SA3	126,788	M	498.12	496.65	485.06	485.18
		S.D.	115.80	115.27	112.67	113.37
SA3	253,354	M	487.86	489.06	480.93	477.80
		S.D.	119.15	121.58	112.99	112.85

when compared to the average linear raw to scaled transformation, than did the average IRT pre-equating results from the previous study. Once again, the discrepancies between the average IRT pre-equating results and the average linear criterion results are greatest in the upper part of the raw score scale.

Conclusions drawn from Figure 5 and 6 are further borne out by the data presented in Table 3. The scaled score summary statistics resulting from application of the current IRT pre-equating restuls are even more discrepant from the intact final form IRT and linear summary statistics than are the summary statistics from the previous study IRT pre-equating results. Hence, as with the previous study, the Form 3ASA3 IRT pre-equating results appear unsatisfactory.

From a review of the average equatings and average residual plots for Form 3BSA3 in Figure 7, somewhat different results from those for Form 3ASA3 can be observed. The average IRT pre-equating from the current study provides, for most of the raw score range, slightly less discrepant results than the average IRT pre-equating from the previous study. Once again, the discrepancies between the average criterion and the average IRT pre-equating results from the current study are, for the most part, in exactly the same direction as the discrepancies for the average pre-equating results from the previous study; they are just slightly less extreme for most of the raw score scale greater than zero. Both IRT pre-equatings provide higher raw to scaled conversion lines than that provided by the intact final form IRT criterion equating through the upper part of the raw score scale. The discrepancy between the current average IRT pre-equating results and the

intact final form criterion results is greatest around raw formula scores of 40 to 50 and in this region the discrepancy is, as was the case for 3ASA3, between 15 and 20 scaled score points.

Using the average linear equating results actually used operationally to place Form 3BSA3 on the 200 to 800 score reporting scale as a criterion (see Figure 8), the results are completely consistent with those in Figure 7; the current average IRT pre-equating provides slightly less discrepant results than the average IRT pre-equating from the previous study. Once again, the discrepancies between the average pre-equating results and the average linear criterion results are greatest in the upper part of the raw score scale.

Conclusions drawn from Figures 7 and 8 are corroborated by the data presented in Table 3. The scaled score summary statistics resulting from application of the current IRT pre-equating results are somewhat closer to the intact final form IRT and linear summary statistics than are the summary statistics from the previous study IRT pre-equating results. In sum, the data suggests that the IRT pre-equating results from the current study provide an improvement over the IRT pre-equating results from the previous study. Unfortunately, the improvement is only slight, and with maximum scaled score differences of upwards of 15 points or greater between the current IRT pre-equating and the criterion equatings, the current pre-equating results must still be deemed unacceptable.

Supplemental Equating Results

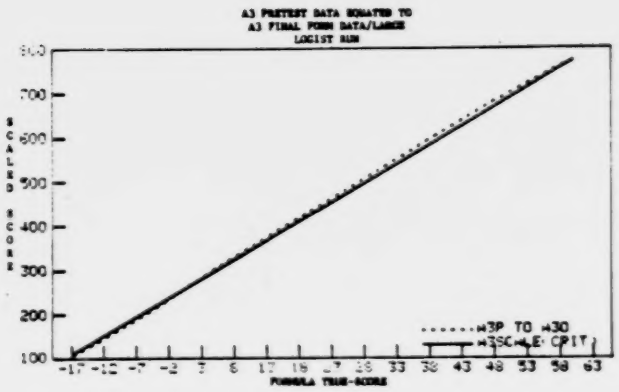
When the data from the previous study were run in the large LOGIST run for this study, 3ASA3 and 3BSA3 data from administration of the forms in intact final form fashion were also included in the calibration. This provides two sets of parameter estimates for each of the items in 3ASA3 and two sets for 3BSA3; one set of parameter estimates are based on the results of administering the 3ASA3 (or 3BSA3) items in a variety of pretests and the other set of parameter estimates are based on the intact final form administration. Further, there is no need to link these sets of parameter estimates in order to make comparisons, as was the case in the previous study; they are automatically on the same scale because they were included in the same LOGIST run. Form 3ASA3 can be equated to itself, as can Form 3BSA3. If nothing is aberrant about either set of parameter estimates, then aside from estimation error, this equating should result in an identity transformation.

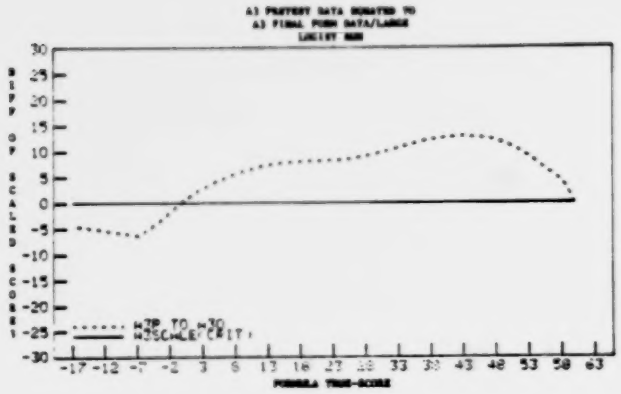
Figure 9 contains equating and residual plots for Forms 3ASA3 and 3BSA3 expressed on the scaled score metric. The criterion transformation is simply the linear raw to scale transformation used to place the form on scale the first time it was administered operationally as an intact form. The other transformation is the result of equating 3ASA3 (or 3BSA3) based on pretest parameter estimates to 3ASA3 (3BSA3) based on final form parameter estimates and then using this transformation in conjunction with the final form linear raw to scale transformation to derive a new raw to scaled transformation. To the extent that the sets of parameter estimates are different, this will result in a different raw to scaled transformation from the linear one.

Equating Plot

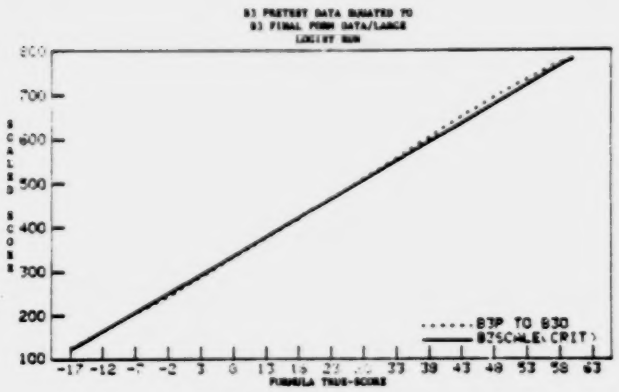
Residual Plot

3ASA3





3BSA3



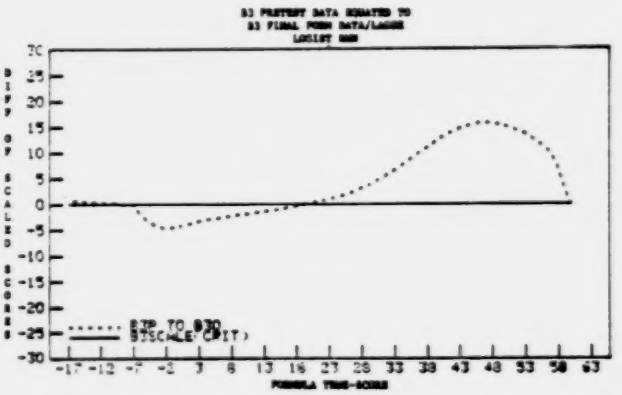
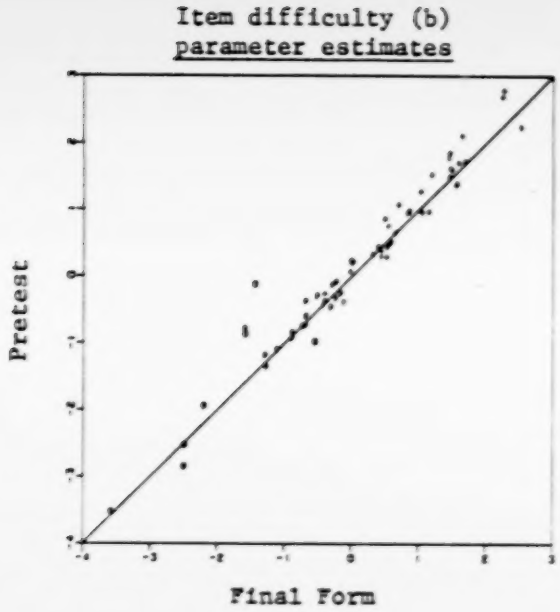
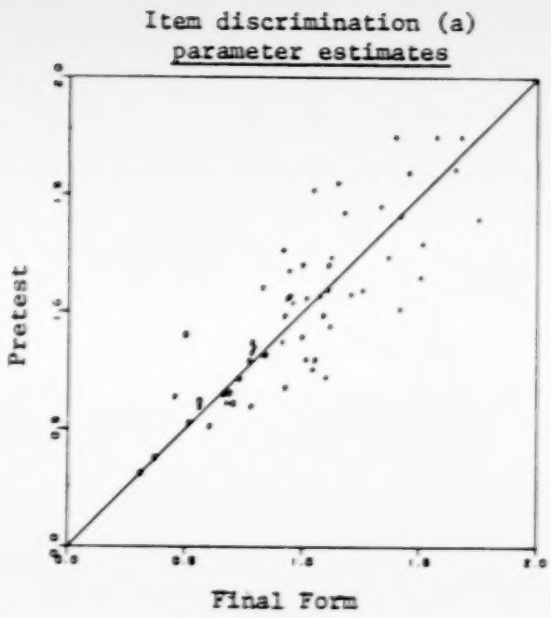


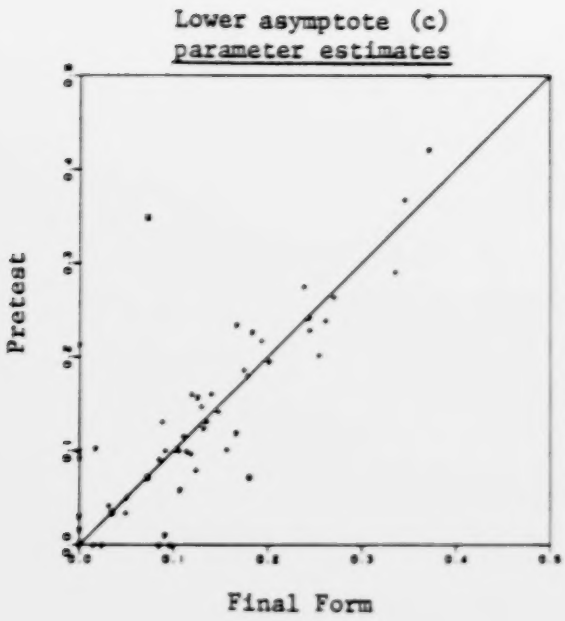
Figure 9: SAT-math Forms 3ASA3 and 3BSA3 - Plots of 1) linear criterion raw to scaled transformation for 3ASA3 (3BSA3) compared to raw to scaled transformation resulting from equating 3ASA3 (3BSA3) to itself using item parameter estimates from the large LOGIST run generated from administration of the 3ASA3 (3BSA3) items in pretest and intact final form fashion, and 2) differences between scaled scores (linear criterion equating minus equating resulting from equating 3ASA3 (3BSA3) to itself) resulting from the equatings. Raw to scaled score transformations were produced, rather than raw score to raw score transformations, so that the equating and residual plots would present data on scales comparable to these in Figures 3-8.

As can be seen in Figure 9, the two raw to scaled transformations are quite different. Equating 3ASA3 (or 3BSA3) to itself through use of the pretest and final form item parameter estimates results in a raw to scaled transformation that is higher through most of the upper part of the raw score scale. It should be noted that the plots in Figure 9 are completely consistent in appearance with the average plots contained in Figures 5 and 7; they are also consistent with the plots from the Eignor (1985) study. The conclusion to be drawn here must be the same as that drawn in the previous study. The higher raw to scaled transformation has to result from the fact that certain of the 3ASA3 and 3BSA3 items have item difficulty parameter estimates that make them appear to be more difficult when given in pretest than in final form.

To corroborate this conclusion, two way plots of pretest versus final form item parameter estimates from the large LOGIST run were prepared. The plots for 3ASA3 are contained in Figure 10 while comparable plots for 3BSA3 are contained in Figure 11. The plots of the pretest and final form item difficulty estimates in Figures 10 and 11 are indeed consistent with the above conclusion. There are a larger number of individual points lying above the diagonal than below; this is also indicated in the small table in Figures 10 and 11. Points lying on the diagonal are items that have no difference between pretest and final form difficulty parameter estimates. Points above the diagonal indicate items that were estimated to be more difficult in pretest than in the final form. Two-way plots of item discrimination and lower asymptote parameter estimates in Figures 10 and 11, while indicating a good deal more variability in individual item parameter





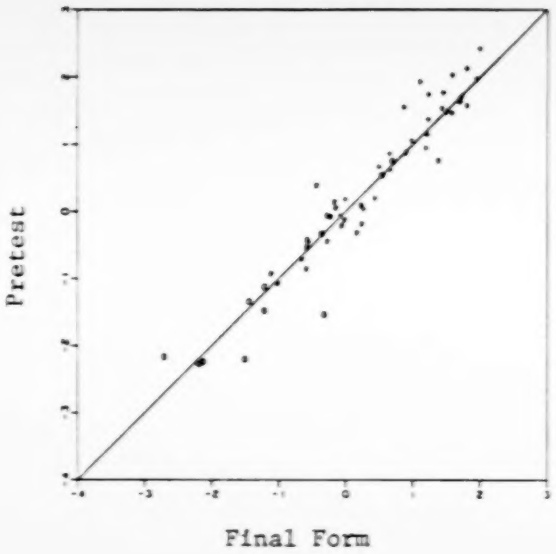


Points above the main diagonal

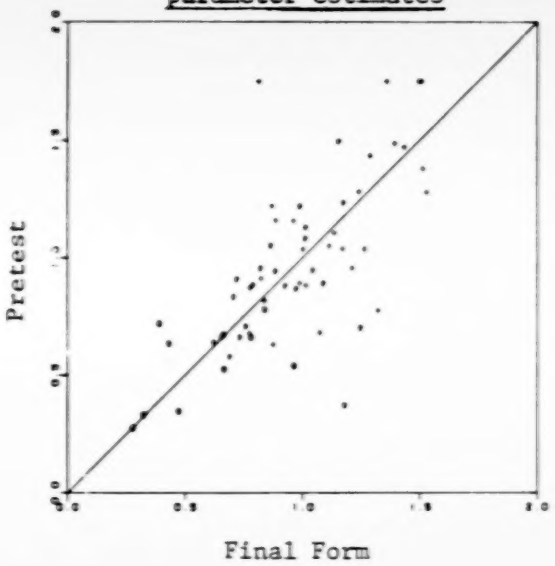
Parameter estimate	Number	Percentage
b	40	67
a	29	48
c	28	47

Figure 10: Two-way plots of pretest and final form parameter estimates for the 60 3ASA3 mathematics items from the large LOGIST run. Number and percentage of points above the main diagonal in each plot.

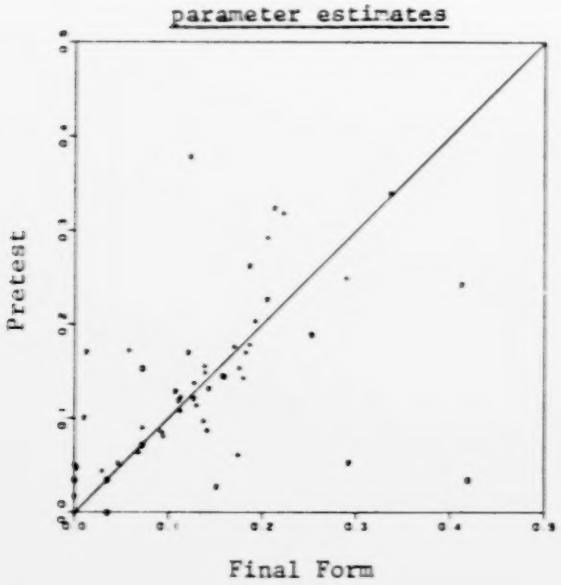
Item difficulty (b)
parameter estimates



Item discrimination (a)
parameter estimates



Lower asumptote (c)
parameter estimates



Points above the main diagonal

Parameter estimate	Number	Percentage
b	36	60
a	31	52
c	28	47

Figure 11: Two-way plots of pretest and final form parameter estimates for the 60 3BSA3 mathematics items from the large LOGIST run. Number and percentage of points above the main diagonal in each plot.

estimates than the two-way difficulty plots, also demonstrate more or less the expected balance of points above and below each diagonal.

Conclusions

The IRT pre-equatings from the second phase of this study provided unexpected results; these results were quite consistent with the results from the previous study. The authors of this study had expected that either multiple usage of the parameter scaling procedure (Stocking and Lord, 1983) or the lack of adequate cross-links in the previous calibration design were responsible for the fact that many of the items were estimated as being more difficult in pretest than in final form and for the fact that the IRT pre-equating results were disparate from the intact final form equating results in that study. The poor pre-equating results from this study indicate that neither can be used as an explanation for the poor IRT pre-equating results from the previous study. It must be concluded that the three parameter logistic model, as implemented by LOGIST, cannot successfully handle the specific SAT-mathematical data used in this study. The problem must lie either in the data or in the calibration process.

A number of possible explanations offered by Eignor (1985) for the poor pre-equating results from the previous study are still relevant. The first three potential explanations were seen as less likely contributors to the poor pre-equating results from the previous study. Given the results of the current study, the likelihood of their providing an explanation for the poor pre-equating results has increased. The first two possibilities are, unfortunately, very difficult to isolate or investigate further. They are:

1. In this study and the previous study, only the pretest items in pretest sections that were needed to perform the actual pre-equatings were calibrated. This seemed a reasonable thing to do in the previous study; the expectation was that this would possibly even improve the calibration process. A certain number of the other pretest items in the various pretest sections were found to be faulty, and these items would certainly have caused problems in estimation if they were included. However, it still seems reasonable to question whether the difficulty estimates for the pretest items would have been different had the entire pretest sections been calibrated. The authors considered including entire pretest sections in the large LOGIST calibration run, but this would have increased the total number of items from 1,600 to approximately 2,325. Given the size of the LOGIST run without the additional data and the potential for problems in getting the LOGIST calibration procedure to converge with such massive amounts of data, it was decided to forego investigating this further. Thus, it remains as a potential, though improbable, explanation for the poor results in this and the previous study.
2. The discrepancies in the pretest and final form item difficulty estimates, and the resultant IRT equatings, may be due to context effects (i.e., the relationship between the item of interest and adjacent items), which because of the nature of the design of this and the previous study cannot readily be isolated. While it is reasonable to assume that the context in which an item occurs may

affect the parameter estimates that result (see Yen, 1980), it is a bit more difficult to envision that these context effects would be predominately in the same direction, which would have to have been the case, at least in terms of item difficulty parameter estimates, in both studies. Also, a careful review of all items, both in pretest and final form, that were identified as having widely discrepant item response functions in the previous study failed to locate any sort of readily apparent context effect.

3. The discrepancies in the pretest and final form item difficulty estimates, and the resultant IRT equatings, may be the result of differences in the ability levels of the groups used for calibration purposes. Theoretically, IRT item parameters are supposed to be independent of the ability level of the group used in the calibration process; in practice, this is not always the case, in particular for item difficulty estimates (Cook, Eignor, and Petersen, 1982). Eignor (1985) provided scaled score summary data for the samples taking SAT-mathematical Form 3ASA3 and Form 3BSA3 items in pretest and intact final form fashion. This data clearly indicated that the above hypothesis warranted further investigation. For Form 3ASA3, 92.9% of the samples taking the items in pretest fashion had lower scaled score means than the sample taking the items in intact final form fashion; for Form 3BSA3, this figure was 59.1%. In a sequel to this study (Stocking and Eignor, 1985), the authors will investigate this hypotheses further via data simulation procedures. Using Form

3ASA3 and equating section fn, as calibrated in LOGIST run number 8 depicted in Figure 1, and treating 3ASA3 item and ability parameter estimates as true parameters, a number of samples will be created whose ability distributions differ in a systematic fashion from the "true" 3ASA3 ability distribution. Using these ability distributions and the "true" item parameters, item response data for the 3ASA3 items will be simulated in each sample and then calibrated together in one concurrent LOGIST run using equating section fn as the common set of items across all samples. Using item parameter estimates generated in each sample, 3ASA3 will then be equated to itself a number of times. Differences among the equatings should provide a clear indication of how differences in ability distributions can effect equating results.

4. Finally, one other potential explanation for the poor pre-equating results from this study has recently been offered. It, is at present only a hypothesis, and would require further investigation. Results from usage of the concurrent calibration design for the operational IRT equating of SAT final forms have provided an indication that the characteristics of the common items used to provide internal linkages in the concurrent design can affect the quality of the resulting equatings. It is quite possible that items may have contributed to problematic item parameter scalings in the previous study are also causing problematic internal linkages in the large concurrent LOGIST run. Individual items in the item parameter scalings from the previous study have not been carefully studied to date; only the overall quality of the scalings were ascertained

and found to be acceptable. Revisiting the items in these scalings, removing poorly performing items, and redoing the scalings might possibly improve on the pre-equating results from the previous study.

In summary, a study of the item parameter scalings would seem to be an important topic to pursue if the planned investigation of the possible effects of the ability levels of the calibration samples on parameter estimates does not provide an explanation for the poor pre-equating results. However, if the results of both of these planned studies do not provide explanations for the results of this and the previous pre-equating study, then it will be reasonable to conclude that pre-equating is not a viable procedure for placing new forms of the SAT on scale.

References

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Bejar, I. I., and Wingersky, M. S. A study of pre-equating based on item response theory. Applied Psychological Measurement, 1982, 6, 309-325.

Cook, L. L., and Eignor, D. R. Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Cook, L. L., Eignor, D. R., and Petersen, N. S. A study of the temporal stability of IRT item parameter estimates. A paper presented at the annual meeting of AERA, New York, 1982.

Cook, L. L., McHale, F. J., Eignor, D. R., Petersen, N. S., and Dorans, N. J. Item response theory equating of aptitude tests: A partial pre-calibration design. Paper presented at the annual meeting of AERA, Chicago, 1985.

Eignor, D. R. An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections. ETS Research Report 85-10. Princeton, NJ: Educational Testing Service, 1985.

Eignor, D. R., and Cook, L. L. A study of the stability of IRT parameter estimates between pretest and final form. Paper presented at the annual meeting of NCME, New Orleans, 1984.

Kingston, N. M., and Dorans, N. J. The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. ETS Research Report 82-12. Princeton, NJ: Educational Testing Service, 1982.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980

Petersen, N. S., Cook, L. L., and Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 1983, 8, 137-156.

Stocking, M. L. Documentation for general IRT equating program. Unpublished report. Princeton, NJ: Educational Testing Service, 1981.

Stocking, M. L., and Eignor, D. R. The impact of different ability distributions on IRT pre-equating. A paper to be prepared for the ETS Program Research Planning Council, 1985.

Stocking, M. L., and Lord, F. M. Developing a common metric in item response theory. Applied Psychological Measurement, 1983, 7, 201-210.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Wingersky, M. S., Barton, M. A., and Lord, F. M. LOGIST V user's guide. Princeton, NJ: Educational Testing Service, 1982.

Yen, W. M. The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 1980, 17, 297-311.